W214–W219 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki385

GlyProt: *in silico* glycosylation of proteins

Andreas Bohne-Lang* and Claus-Wilhelm von der Lieth

German Cancer Research Center Heidelberg, Central Spectroscopy–Molecular Modeling,
Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Received February 11, 2005; Revised and Accepted March 9, 2005

ABSTRACT

GlyProt (http://www.glycosciences.de/glyprot/) is a web-based tool that enables meaningful *N*-glycan conformations to be attached to all the spatially accessible potential N-glycosylation sites of a known three-dimensional (3D) protein structure. The probabilities of physicochemical properties such as mass, accessible surface and radius of gyration are calculated. The purpose of this service is to provide rapid access to reliable 3D models of glycoproteins, which can subsequently be refined by using more elaborate simulations and validated by comparing the generated models with experimental data.

INTRODUCTION

The human genome appears to encode no more than 25 000 proteins (1). This relatively small number of genes compared with the genome of other species has been one of the big surprises to come out of the Human Genome Project. A major challenge is to understand how post-translational events affect the activities and functions of these proteins in relation to health and disease. Among these, glycosylation is by far the most frequent; more than half of all the proteins in the human body have glycan molecules attached (2,3). Glycosylated proteins are ubiquitous components of extracellular matrices and cellular surfaces. Their oligosaccharide moieties are implicated in a wide range of cell–cell and cell–matrix recognition events. *N*-glycans covalently connected to proteins constitute highly flexible molecules. Therefore, only a small number of glycan structures are available for which sufficient electron density for an entire oligosaccharide chain can be detected (4). Unambiguous structure determination based on NMR-derived geometric constraints alone is often not possible (5). Time-consuming computational approaches such as Monte Carlo calculations and molecular dynamics simulations have been widely used to explore the conformational space accessible to complex carbohydrates (6,7).

For reasons that are not well understood, not all Asn-X-Ser/Thr sequons are glycosylated. Unfortunately, the unambiguous determination of occupied N-glycosylation sites is experimentally demanding and can vary between different cellular locations. The aims of GlyProt are (i) to evaluate whether a potential N-glycosylation site is spatially accessible, (ii) to generate reasonable three-dimensional (3D) models of glycoproteins with user-definable glycan moieties and (iii) to provide some evidence on how the physicochemical parameters can change between the varying glycoforms of a protein.

MATERIALS AND METHODS

The 3D structure of a protein in Protein Data Bank (PDB) format is required as input (see dataflow given in Figure 1). The protein structure can be either taken directly from the PDB or uploaded from a local computer. Potential N-glycosylation sites (sequon: Asn-X-Ser/Thr, where X is not Pro) are automatically detected and highlighted using the one-letter amino acid code. In cases where experimental coordinates with already attached glycans are provided, the internal coordinates (distance between the *N* of the Asn-sidechain and the C1 of the attached β-D-GlcpNAc and the torsion angles determining the orientation of the glycan moiety) are displayed.

Orientation of the *N*-glycans

The orientation of the attached *N*-glycan relative to the glycosylation site is described by the four consecutive torsion angles χ_1, χ_2, Φ and Ψ (for definition see Table 1). It is well known from the analysis of the experimentally available 3D structures of glycoproteins (8,9) that preferred orientations of the glycan moiety relative to the protein exist (Figure 2). The current version of the PDB contains nearly 3000 *N*-glycan chains. Conformational maps indicating the populated areas for all four torsion angles can be easily obtained using the GlyTorsion tool (http://www.glycosciences.de/glytorsion/) from the Carbohydrate Structure Suite (10).

It is assumed that the Man_3 *N*-glycan core exhibits one dominant, relatively rigid conformation. This assumption is supported by the analysis of experimentally determined torsion angles for the corresponding glycosidic linkages in the PDB (Table 2 and Figure 3). Only the 1–6 linkage exhibits two significantly populated conformations, whereas the other three linkages constitute only one highly populated conformation.

*To whom correspondence should be addressed. Tel: +49 6221 42 4541; Fax: +49 6221 42 3669; Email: a.bohne@dkfz-heidelberg.de

To evaluate whether a potential glycosylation site is spatially accessible, a program written in C is used to connect the Man_3 *N*-glycan core to the protein and test all possible angle sets. The frequency of occurrence of the four relevant torsion angles (Table 1) is used to orient the *N*-glycan core. Next, the program evaluates whether atoms of the attached glycan moiety overlap with the protein. If spatial overlaps are detected, the model is rejected and the next most frequently observed orientation of the glycan moiety is applied. Table 1 lists the values of the four relevant torsion angles and the succession in which they are applied. This procedure is repeated until a structure with no or minor overlap has been found. If all orientations listed in Table 1 have been applied and all

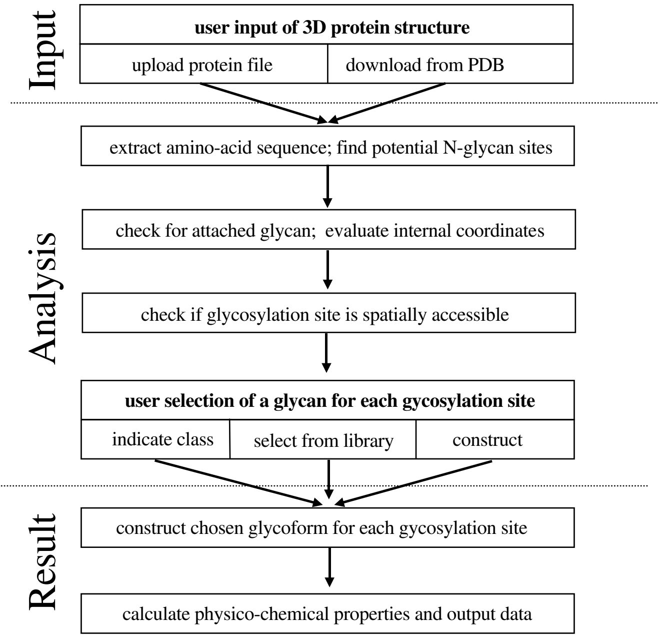

Input
user input of 3D protein structure
upload protein file
download from PDB
Analysis
extract amino-acid sequence; find potential N-glycan sites
check for attached glycan; evaluate internal coordinates
check if glycosylation site is spatially accessible
user selection of a glycan for each gycosylation site
indicate class
select from library
construct
Result
construct chosen glycoform for each gycosylation site
calculate physico-chemical properties and output data

Figure 1. Dataflow of GlyProt.

Table 1. Definition of torsion angles defining the orientation of the glycan moiety relative to the protein and hierarchy of applied torsion angles

Name	Definition of torsion angles	Hierarchy of applied torsion angles
χ_1	$N–C_\alpha–C_\beta–C_\gamma$	180, 200, 300, 280, 60, 80, 40, 220, 320
χ_2	$C_\alpha–C_\beta–C_\gamma–O$	340, 320, 20, 0, 40, 60, 280, 80, 280
Ψ_n	$C_1–N_1–C_\gamma–C_\beta$	160, 180, 200
ϕ_n	$O_5–C_1–N_1–C_\gamma$	260, 280, 240, 220, 300

Table 2. Torsion angles for glycosidic linkages of the *N*-glycan core region

Linkage	Φ	Ψ	*P* (%)	No. in PDB
β-D-GlcpNAc-(1–4)-β-D-GlcpNAc	280	240	85	472
β-D-Manp-(1–4)-β-D-GlcpNAc	280	240	77	1187
α-D-Manp-(1–3)-β-D-Manp	80	100	83	277
α-D-Manp-(1–6)-β-D-Manp	60	80	48	218
	60	160	28	

Analysis of the PDB entries was carried out with the GlyTorsion tool. With $\Phi = O_{\mathrm{ring}} - X_{A\nu} - O_{\gamma\lambda\psi} - X_{\alpha\gamma}$ and $\Psi = X_{A\nu} - O_{\gamma\lambda\psi} - X_{\alpha\gamma} - C_{\alpha g-1}$. Analysis of the PDB entries was carried out with the the GlyTorsion tool.

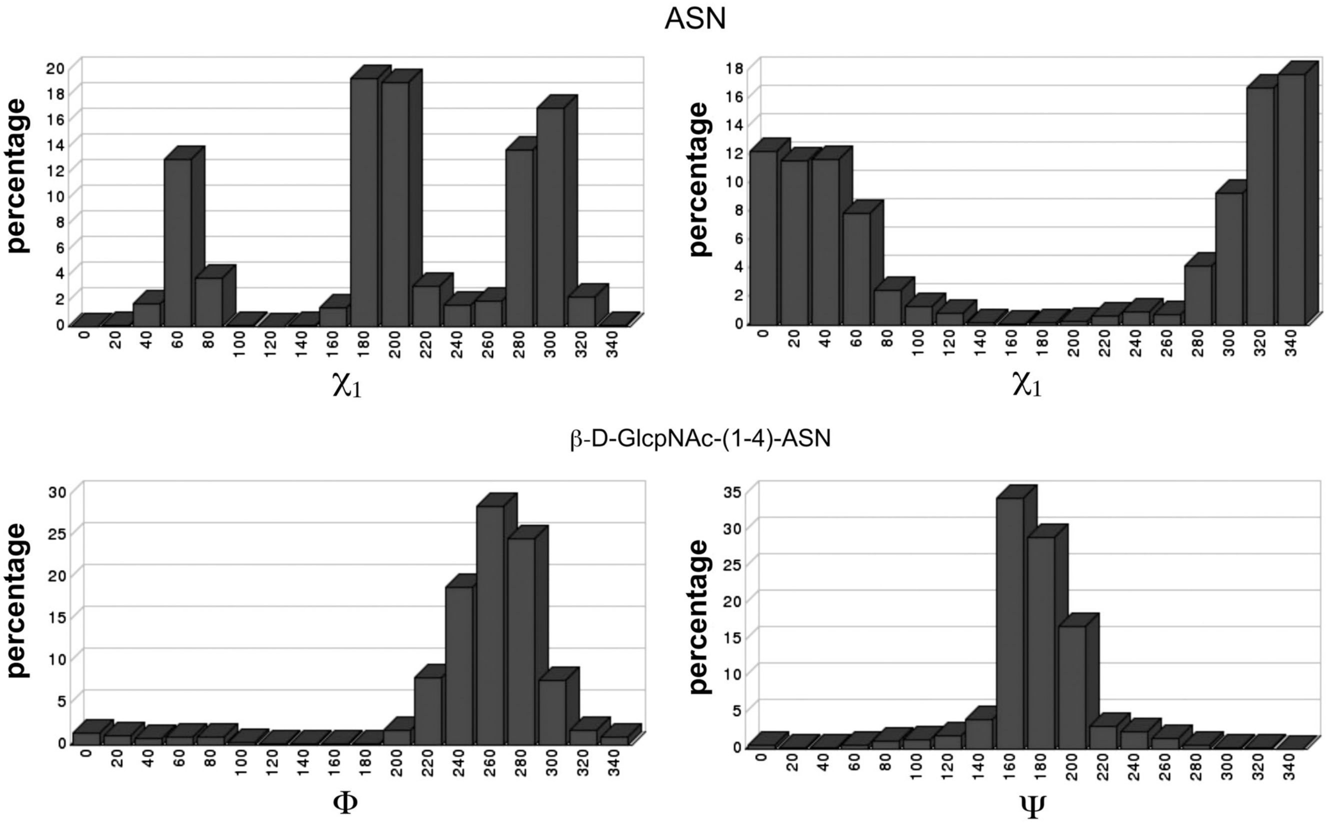


Figure 2. Statistical analysis of the PDB for torsion angles determining the orientation of the glycan moiety relative to the protein.

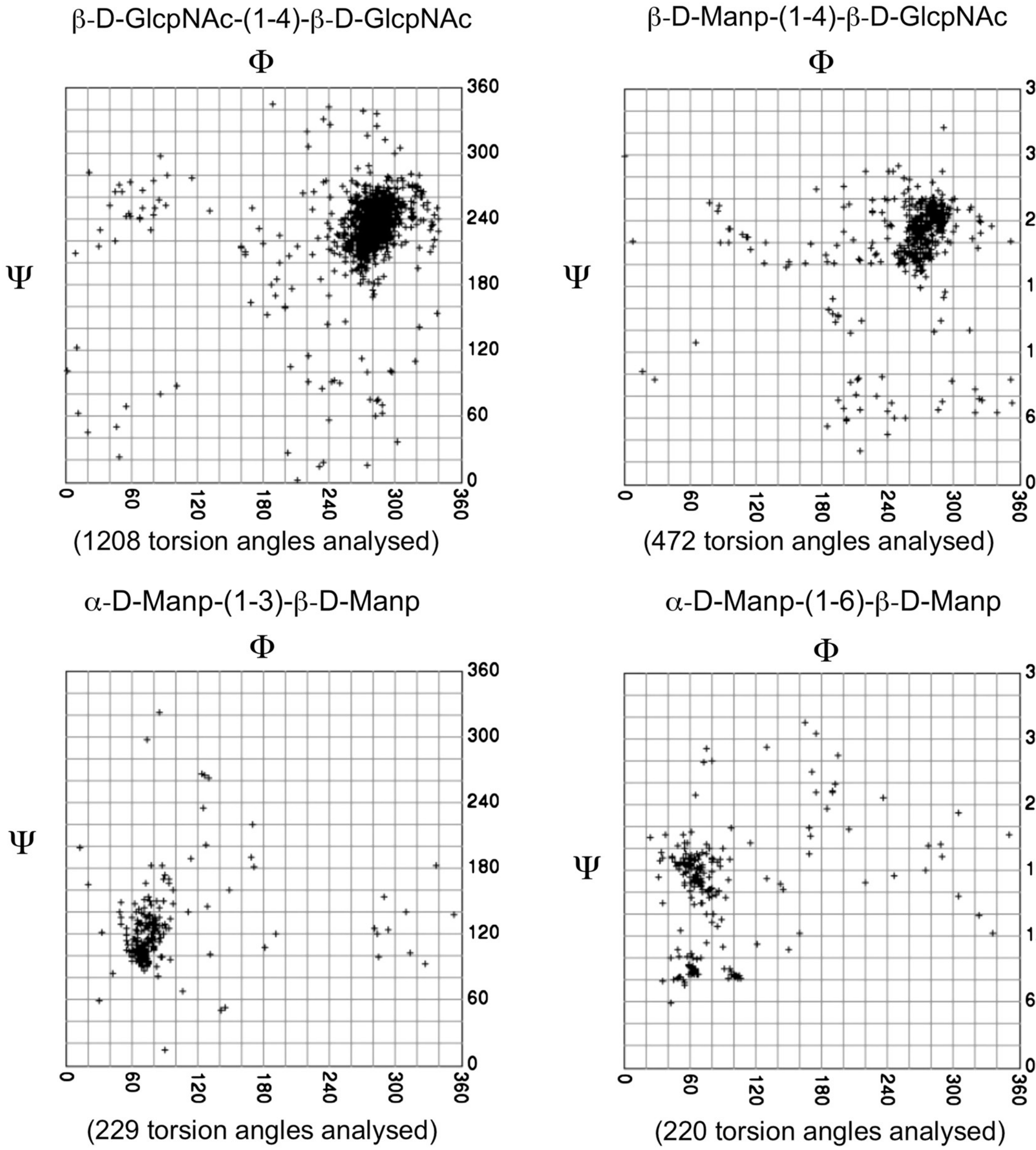


Figure 3. Statistical analysis of the PDB for glycosidic torsion angles determining the conformation of the *N*-glycan core.

resulting glycoprotein structures exhibit overlapping atoms, it is assumed that the glycosylation site is spatially inaccessible and therefore cannot be glycosylated.

Construction of user-definable glycoproteins

For each spatially accessible potential N-glycosylation site three options are offered for selecting the *N*-glycan to be connected. The user can

(i) select the type of *N*-glycan (e.g. oligomannose rich, complex, hybrid, very large); by default a typical structure for each class is taken;

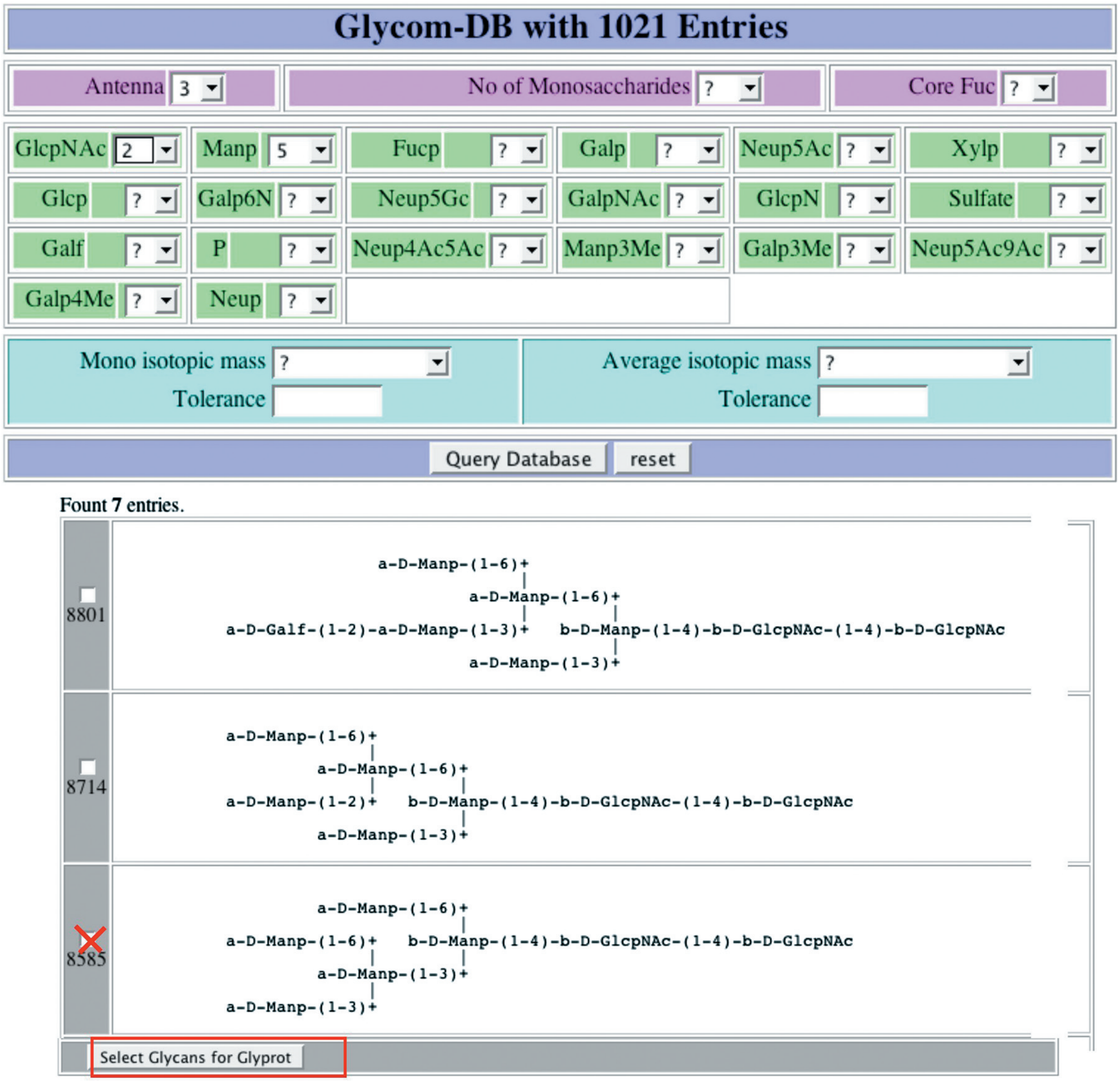
(ii) select an *N*-glycan from a database of >1000 structures (Figure 4) constructed using SWEET-II (11) and optimized using the TINKER MM3 force field (http://dasher.wustl.edu/tinker/); the database is searchable by *N*-glycan composition;

(iii) construct the desired *N*-glycan using SWEET-II by user input of the desired structure using the extended IUPAC nomenclature.

If the coordinates provided already contain attached *N*-glycans, the user can either accept this orientation or use the procedure described above to align the glycan moiety.

RESULTS

The atomic coordinates of the desired glycoprotein are given in PDB format, and they are immediately displayed using the Java applet Jmol (http://jmol.sourceforge.net/). The coordinates can be downloaded and used as input for many 3D visualization programs (see Figure 5). In addition, some physicochemical parameters for the non-glycosylated and the glycosylated protein are displayed to provide a general



Glycom-DB with 1021 Entries

Antenna 3	No of Monosaccharides ?	Core Fuc ?			
GlcpNAc 2	Manp 5	Fucp ?	Galp ?	Neup5Ac ?	Xylp ?
Glcp ?	Galp6N ?	Neup5Gc ?	GalpNAc ?	GlcpN ?	Sulfate ?
Galf ?	P ?	Neup4Ac5Ac ?	Manp3Me ?	Galp3Me ?	Neup5Ac9Ac ?
Galp4Me ?	Neup ?				
Mono isotopic mass ?	Tolerance	Average isotopic mass ?	Tolerance		

Fount 7 entries.

□ 8801

```
               a-D-Manp-(1-6)+
                             |
                        a-D-Manp-(1-6)+
                             |         |
a-D-Galf-(1-2)-a-D-Manp-(1-3)+    b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc
                                       |
                        a-D-Manp-(1-3)+
```

□ 8714

```
a-D-Manp-(1-6)+
              |
         a-D-Manp-(1-6)+
              |        |
a-D-Manp-(1-2)+    b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc
                       |
         a-D-Manp-(1-3)+
```

☒ 8585

```
         a-D-Manp-(1-6)+
                       |
a-D-Manp-(1-6)+    b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc
              |        |
         a-D-Manp-(1-3)+
              |
a-D-Manp-(1-3)+
```

Figure 4. Input spreadsheet (top) used to query the database, which contains >1000 3D structures of *N*-glycans (bottom). The user indicates the desired glycoform by checking the corresponding selection box.

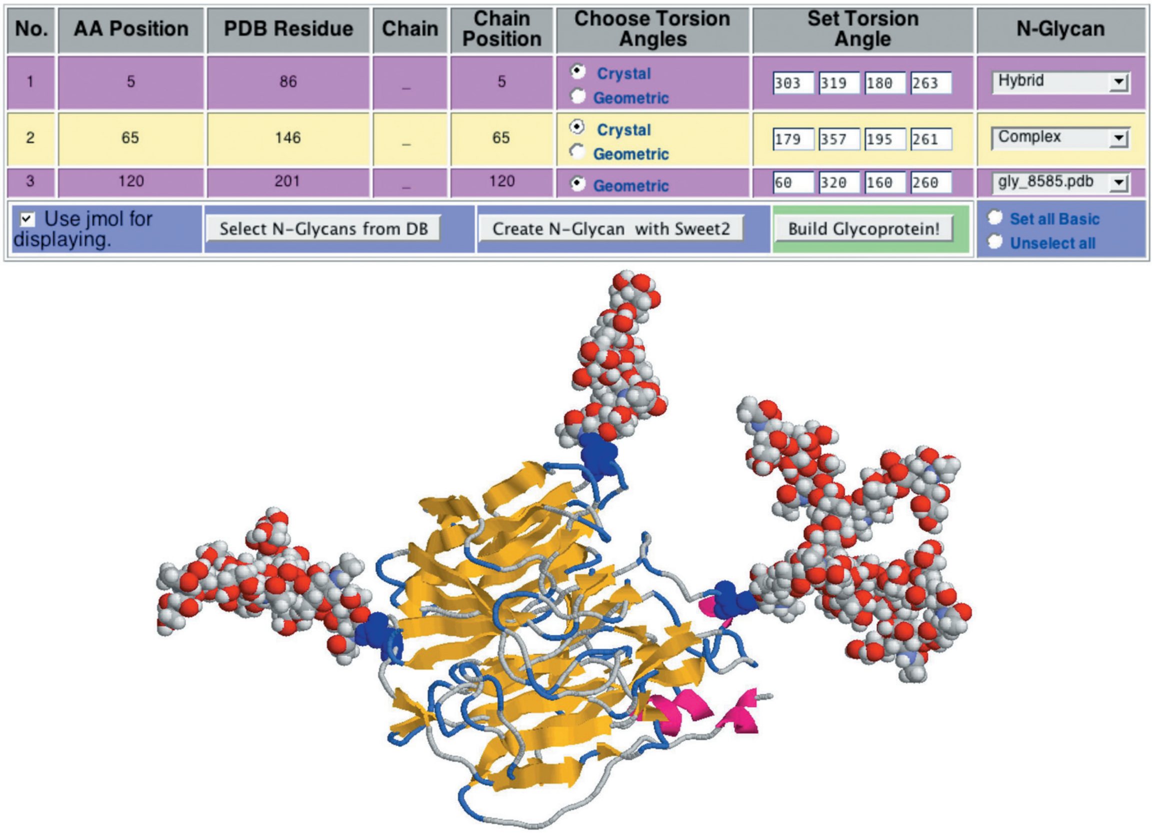
delineation of the changes caused by the selected glycoform (see Table 3). The program Surface Racer (12) is used to calculate the solvent accessible surface of both molecules. The generally observed increase of the polar surface area as a result of glycosylation reflects the well-known experience that glycoproteins exhibit higher solubility.

DISCUSSION

GlyProt enables rapid Internet-based access to reasonable 3D model of glycoproteins. Although it is estimated that >50% of all proteins are glycosylated (2,3), only ~5% of all PDB entries have attached glycan chains (4). Moreover, only a few entries in the PDB contain X-ray diffraction data with sufficient electron density to detect an entire oligosaccharide chain. The 3D models of glycoproteins constructed with GlyProt can provide some evidence on which areas of a protein are captured by a certain glycoform and whether, for example, a binding site is covered so that the biological activity of a protein may be influenced.

Simply because of their large size and hydrophilicity, glycans can alter the physicochemical properties of a glycoprotein, making it more soluble, reducing backbone flexibility and thus leading to increased protein stability, protecting it from proteolysis, and so on. The calculation of some characteristic physicochemical parameters will help in the evaluation and explanation of the varying properties of different glycoforms. Of the therapeutic proteins on the market, ~60% are glycoproteins (13). Often, the removal of *N*-glycans results in a protein with a very short half-life and virtually no activity *in vivo* (13).

A comprehensive evaluation of the impact of varying glycoforms on protein function is hampered by the high conformational flexibility of glycan structures. Based on the statistical analysis of experimentally known glycan conformations, GlyProt constructs a reasonable conformation out of



No.	AA Position	PDB Residue	Chain	Chain Position	Choose Torsion Angles	Set Torsion Angle	N-Glycan
1	5	86	_	5	● Crystal ○ Geometric	303 319 180 263	Hybrid
2	65	146	_	65	● Crystal ○ Geometric	179 357 195 261	Complex
3	120	201	_	120	● Geometric	60 320 160 260	gly_8585.pdb
☑ Use jmol for displaying.	Select N-Glycans from DB			Create N-Glycan with Sweet2		Build Glycoprotein!	○ Set all Basic ○ Unselect all

Figure 5. User interface (top) to select the desired glycoform for each gycosylation site. Visualization (bottom) of the constructed glycoprotein. The protein part is given as a cartoon representation; the glycan part as a spacefill model.

Table 3. Comparison of some characteristic physicochemical properties of the pure Influenza A Subtype N9 Neuraminidase (14) and the constructed glycoform

	Pure protein	Glycoprotein
Number of heavy atoms	3069	3252
Total ASA (Å^2)	15 122.39	18 124.59
Polar ASA (Å^2)	7626.93	9096.65
Non-polar ASA (Å^2)	7495.46	9027.94
+charge ASA (Å^2)	1770.88	1747.45
-charge ASA (Å^2)	1402.60	1402.60
Structure contains cavities	25	25
Radius of gyration (Å)	20.746696	22.44755
Monoisotopc mass (a.m.u.)	43 674.90154	48 297.494696
Average mass (a.m.u.)	43 702.95044	48 328.11854

ASA, solvent accessible surface area (in Å^2).

a manifold. However, a more realistic analysis would require the complete conformational space that is accessible to a glycan at a given glycosylation site to be sacnned. Therefore, we intend to expand the GlyProt service with an option allowing the exploration of the conformational space accessible to an *N*-glycan, which is covalently bound to a specific glycosylation site. A similar approach has already been successfully applied to rapidly generate a representative ensemble of conformations of single *N*-glycan molecules (6). This algorithm is based on a comprehensive set of conformations of *N*-glycan fragments that were derived from molecular dynamics simulations. However, this approach would assume a protein conformation that remains unchanged through the attachment of varying glycans. In order to allow conformational changes of the protein backbone, only force-field-based, time-consuming simulation approaches such as molecular dynamics with inclusion of explicit water molecules would be appropriate.

ACKNOWLEDGEMENTS

The development of GlyProt is funded by a grant from the German Research Council (Deutsche Forschungsgemeinschaft, DFG) within the digital library program. Funding to pay the Open Access publication charges for this article was provided by DFG.

Conflict of interest statement. None declared.

REFERENCES

1. International Human Genome Sequencing Consortium (2004), Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
2. Apweiler,R., Hermjakob,H. and Sharon,N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, **1473**, 4–8.
3. Ben-Dor,S., Esterman,N. and Rubin,E. (2004) Biases and complex patterns in the residues flanking protein N-glycosylation sites. *Glycobiology*, **14**, 95–101.
4. Luetteke,T., Frank,M. and von der Lieth,C.W. (2004) Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr. Res.*, **339**, 1015–1020.

5. Imberty,A. and Perez,S. (2000) Structure, conformation, and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations. *Chem. Rev.*, **100**, 4567–4588.
6. Frank,M., Bohne-Lang,A., Wetter,T. and von der Lieth,C.W. (2002) Rapid generation of a representative ensemble of N-glycan conformations. *In Silico Biol.*, **2**, 427–439.
7. Woods,R.J. (1998) Computational carbohydrate chemistry: what theoretical methods can tell us. *Glycoconj. J.*, **15**, 209–216.
8. Imberty,A. and Perez,S. (1995) Stereochemistry of the N-glycosylation sites in glycoproteins. *Protein Eng.*, **8**, 699–709.
9. Petrescu,A.J., Milac,A.L., Petrescu,S.M., Dwek,R.A. and Wormald,M.R. (2004) Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology*, **14**, 103–114.
10. Lutteke,T., Frank,M. and von der Lieth,C.W. (2005) Carbohydrate structure suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res.*, **33**, D242–D246.
11. Bohne,A., Lang,E. and von der Lieth,C.W. (1998) W3-SWEET: carbohhydrate modeling by internet. *J. Mol. Model.*, **4**, 33–43.
12. Tsodikov,O.V., Record,M.T.,Jr and Sergeev,Y.V. (2002) A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.*, **23**, 600–609.
13. Gerngross,T.U. (2004) Advances in the production of human therapeutic proteins in yeasts and filamentous fungi. *Nat. Biotechnol.*, **22**, 1409–1414.
14. White,C.L., Janakiraman,M.N., Laver,W.G., Philippon,C., Vasella,A., Air,G.M. and Luo,M. (1995) A sialic acid-derived phosphonate analog inhibits different strains of influenza virus neuraminidase with different efficiencies. *J. Mol. Biol.*, **245**, 623–634.